J Mol Model (2012) 18:1917–1925
DOI 10.1007/s00894-011-1212-3

ORIGINAL PAPER

Structure-based functional inference of hypothetical proteins from *Mycoplasma hyopneumoniae*

Marbella Maria da Fonsêca · Arnaldo Zaha · Ernesto R. Caffarena · Ana Tereza Ribeiro Vasconcelos

Received: 20 June 2011 / Accepted: 5 August 2011 / Published online: 26 August 2011

Abstract Enzootic pneumonia caused by *Mycoplasma hyopneumoniae* is a major constraint to efficient pork production throughout the world. This pathogen has a small genome with 716 coding sequences, of which 418 are homologous to proteins with known functions. However, almost 42% of the 716 coding sequences are annotated as hypothetical proteins. Alternative methodologies such as threading and comparative modeling can be used to predict structures and functions of such hypothetical proteins. Often, these alternative methods can answer questions about the properties of a model system faster than experiments. In this study, we predicted the structures of seven proteins annotated as hypothetical in *M. hyopneumoniae*, using the structure-based approaches mentioned above. Three proteins were predicted to be involved in metabolic processes, two proteins in transcription and two proteins where no function could be assigned. However, the modeled structures of the last two proteins suggested experimental designs to identify their functions. Our findings are important in diminishing the gap between the lack of annotation of important metabolic pathways and the great number of hypothetical proteins in the *M. hyopneumoniae* genome.

Keywords Comparative modeling · Known function · Modeller · Mollicutes · Threading

M. M. da Fonsêca
Universidade Federal do Rio de Janeiro,
Rio de Janeiro, RJ, Brazil

A. Zaha
Laboratório de Genômica Estrutural e Funcional,
Centro de Biotecnologia, UFRGS,
Porto Alegre, RS, Brazil

E. R. Caffarena
Programa de Computação Científica, Fundação Oswaldo Cruz,
Rio de Janeiro, RJ, Brazil

M. M. da Fonsêca · A. T. R. Vasconcelos (✉)
Laboratório Nacional de Computação Científica,
Laboratório de Bioinformática,
Petrópolis 25651-075 RJ, Brazil
e-mail: atrv@lncc.br

Introduction

Mycoplasmas belong to the class Mollicutes and number approximately 200 species, among which are obligate parasites of humans and commercially important mammals [1] such as pigs. Mycoplasmas are wall-less bacteria distinguished by small genomes of low G+C content. The parasitism, the reduced genome, and the close association of these bacteria with their hosts have contributed to the absence of enzymes involved in important biosynthetic pathways in mycoplasma [2].

Enzootic pneumonia caused by *Mycoplasma hyopneumoniae* is a major constraint to efficient pork production worldwide. The *M. hyopneumoniae* genome contains 920,079 base pairs and 716 protein-coding genes, of which 418 encode proteins that are homologous to proteins with known functions. Currently, there are nearly 1,500 complete genome sequences in GenBank, and half of all of the predicted genes encode proteins having no inferable functions. Similarly, almost 42 % of predicted *M. hyopneumoniae* genes correspond to proteins annotated as hypothetical [3]. This lack of annotation is a particularly intriguing and unsolved issue because, as mentioned above, components of important and essential metabolic pathways present in other organisms have not been identified in mycoplasmas [4, 5].

The BLAST program [6] has contributed significantly to the analysis of nucleotide and amino acid sequences, allowing the prediction of biological functions and evolutionary relationships of genes and proteins [7]. However, this tool can be used with a high degree of confidence only when the sequences are evolutionarily close to each other and the identity between them is over 50%. To overcome these limitations, alternative methodologies such as threading and homology modeling have been used to answer questions about protein properties. These methods are possible because biological processes such as gene duplication and evolutionary divergence occur in many distantly related organisms [8], giving rise to structurally and functionally similar families of proteins. When one or more proteins in a family have experimentally determined structures, it is feasible to model the structures of many other members with reasonable accuracy. This condition is particularly true when the sequence identity between protein domains is ≥30% and larger than 100 residues.

Threading and homology modeling can identify domains and active sites, aiding in placing their locations within a 3D structure (i.e.,surface or buried). Because the determination of a crystal structure is an arduous and sometimes impractical task for some proteins, the homology modeling methodology is a helpful approach that can guide further experimental assays to investigate protein function [9–11]. The rapid growth of structural genomics is producing a considerable number of templates that can be used for homology modeling. The availability of more templates increases the quality of new models, thereby diminishing the gap between computationally derived models and experimental outcomes.

Thus far, mycoplasma genome sequences have not been annotated for activities related to the utilization of ATP, NAD and NADH and amino acid synthesis derived from pyruvate. However, genes corresponding to these activities must exist, otherwise their enzymatic activities would not have been found [12]. This discrepancy suggests that sequence-based methodologies for identifying protein function may not be suitable for mycoplasmas in some cases.

In this study, using structure-based approaches, we were able to predict the function of seven proteins annotated as hypothetical in the *M. hyopneumoniae* genome. Three of the proteins are involved in metabolic processes, a finding that may enhance further studies concerning the metabolism of this bacterium. Another two proteins are involved in transcription, controlling gene expression based on cellular or environmental signals, an important characteristic of pathogenic bacteria such as *M. hyopneumoniae*. Functions for the other two proteins could not be assigned, but their modeled structures suggest experimental designs, which will allow future investigation concerning their function.

Materials and methods

The sequences of 298 proteins belonging to *M. hyopneumoniae* strain 7448, currently annotated as hypothetical in the Genesul database (http://www.genesul.lncc.br/finalMP/), were submitted to two threading programs, GenThreader [13] and Prospect-PSPP [14]. Additionally, these data were analyzed by InterProScan [15] and COG [16], and the functional predictions of these four programs were compared. Thirty-four sequences with the same functional predictions given by at least two of the mentioned programs were selected for manual analysis, resulting in the further selection of seven targets for structural investigation. Firstly, the sequences of these seven proteins were submitted to a PSI-BLAST search at http://blast.ncbi.nlm.nih.gov/Blast.cgi against the Protein Data Bank (PDB). To guide the functional inference of uncharacterized proteins, other bioinformatics tools were used as described elsewhere [17]. These other tools suggested scans against sequence pattern, domain, and family classification databases, as well as structural family databases, to identify conserved, functional residues and to extract homologs for post-hoc comparative modeling.

The local alignment between sequences of the seven selected proteins and their templates provided by threading results was performed using the EMBL/EBI software MAFFT [18] with little manual editing. Sequences were retrieved from NCBI and GeneSul. The BLOSUM30 matrix was used with gap and extension penalties of 1.0 and 0.123, respectively. Afterward, the alignment was used to model the selected proteins with the Modeller program [19] (version 9v8). The overall geometric and stereochemical qualities of the structures were assessed using PROCHECK through the PDBsum server [20] and PROSA-web [21] and are listed in Table 1.

Results and discussion

Threading is based on sequence-to-structure alignment. The target sequence is "threaded" through each template present in databases that contain all known protein folds. Threading is performed by using measures for fitness for each type of amino acid in local structural environments and defined in terms of solvent accessibility and protein secondary structure. If a sequence fits well with a given fold, conserved residues are likely shared suggesting similar functions [22].

The PROSPECT-PSPP threading pipeline showed that 27 (9.06%) of 298 target proteins gave PSI-BLAST hits against the PDB with an E-value<0.0001, indicating the existence of homologs. Additionally, 83 (27.85%) of the proteins had hits against PDB with a Z-score >20, indicating that the fold recognition confidence level was >99%; the remainder of the

Table 1 Sequence and structure information of the selected proteins and their templates

Protein ID	Template[a]	Identity	Evaluation		Proposed function
			PROSA[b]	Ramachandran[c]	
YP_287866			−7.6	97.9 %	Nicotinic acid mononucleotide adenylyltransferase
	2H29	35 %	−8.66		
	2O08	23 %	−7.28		Putative metal-dependent phosphohydrolase (HD domain)
	2OGI	24 %	−8.03		
YP_287786			−6.17	97 %	Nicotinic acid Phosphoribosyltransferase (NAPRTase)
	1YTK	25 %	−8,93		
YP_287675			−7.22	96.5 %	Flavin Adenine Dinucleotide (FAD) synthetase
	1S4M	20 %	−8.68		
YP_287559			−4.52	95.1 %	Participates in the antitermination process (NusB)
	1EY1	21 %	−4.59		
	2JR0	23 %	−5.45		
	1TZT	23 %	−6.52		
	1Q8C	15 %	−6.75		
	1EYV	18 %	−4.6		
YP_288024			−4.22	96.2 %	Key regulator of bacterial transcription initiation (SigE, Sigma-28)
	2Z2S	18 %	−6.33		
	1RP3	20 %	−7.73		
YP_287971			−5.92	100 %	Unknown function. Likely binds to nucleic acids (YlxR)
	1G2R	29 %	−5.47		
YP_288034			−5.9	96.6 %	Unknown function. Likely binds to nucleic acids (YrdC)
	1HRU	18 %	−7.4		

[a] PDB ID

[b] Favored and allowed regions

[c] Z- score templates

proteins had hits with confidence levels between 85 and 99%. The GenThreader results had high confidence levels (certain) for 84 (32.43%) of 259 proteins (total number of hypothetical sequences available in 2005). Detailed information analysis obtained by threading provided interesting and consistent results, which helped us to select seven proteins having the same prediction by the both mentioned programs. In addition, we followed the protocol suggested by Mazumder and Vasudevan [17], as mentioned in Materials and methods. The results proposed homologs with 3D structures available, thereby providing new knowledge to be applied for comparative modeling.

In the following sections, we will discuss the 3D structures and functions predicted for the seven proteins (YP_287866, YP_287786, YP_287675, YP_287559, YP_288024, YP_287971 and YP_288034). Table 1 lists the templates used to obtain the 3D structures and information about the selected protein models.

Completing the NAD biosynthesis pathway

The 3D structure of hypothetical protein YP_287866 exhibits similarity to portions of two different proteins, i.e., the N-terminal region of nicotinate-nucleotide adenylyltransferase (NadD) and the C-terminal region of an uncharacterized histidine-aspartate (HD) domain. Although the steps in NAD biosynthesis and recycling can vary between species, the enzymes involved in these pathways are generally the following: 1) nicotinate phosphoribosyltransferase (NAPRTase) (EC 2.4.2.11), 2) nicotinate mononucleotide adenylyltransferase (NaMNAT or NadD) (EC 2.7.7.1), and 3) NAD synthetase (NadE) (EC 6.3.1.5) (Fig. 1). These enzymes are encoded by the conserved genes *pncB*, *nadD* and *nadE*, respectively. Enzymes involved in NAD biosynthesis have been considered as promising drug targets because they are essential for the viability of most bacteria [23, 24]; however, only *nadE* is annotated in *M. hyopneumoniae*. Because NadD is likely essential, characterization of this enzyme using a structure-based approach for *M. hyopneumoniae* will improve its annotation and add this enzyme to the list of potential therapeutic targets.

The sequence similarity between the YP_287866 N-terminal region and other nicotinate-nucleotide adenylyltransferases is low (approximately 30%); however, the proteins share two highly conserved ATP-binding motifs,

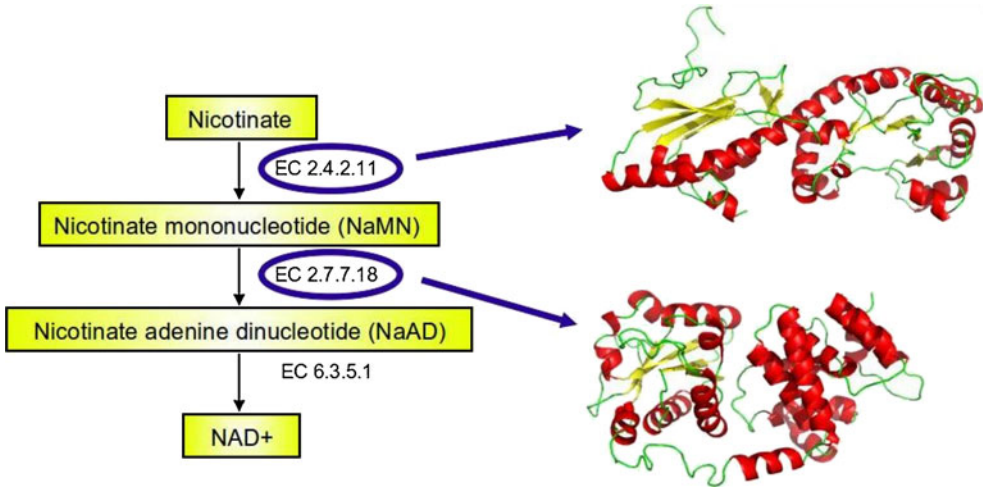

Fig. 1 Simplified NAD biosynthesis pathway proposed for *M. hyopneumoniae*. Highlighted in blue circles are the EC numbers of the enzymes whose 3D structure was predicted in this study. YP_287786 is proposed to be EC 2.4.2.11, a nicotinate phosphoribosyltransferase. YP_287866 (N-terminal region) is suggested to be a nicotinate-nucleotide adenylyltransferase, EC 2.7.7.18. EC 6.3.5.1 is the enzyme NadE, already annotated in *M. hyopneumoniae*. The 3D structures were obtained using comparative modeling methodology, and the structures were rendered with Pymol (www.pymol.org)

GXXXPX(T/H)XX and SX(T/S)XXR. The crystal structures of many NaMNAT proteins [25–29] reveal the residues involved in their function, such as the following: 1) His20, Ser162, Arg167 and the essential His17 in the enzymes from *Pseudomonas aeruginosa* [30], *Escherichia coli* [31] and *B. subtilis* [28], located in the ATP binding site, 2) Thr87 and Trp117 that interact with the substrate nicotinic acidyl, and 3) Arg134 that interacts with the adenosine.

The template selected to obtain the 3D structure of the YP_287866 N-terminal region was the crystal structure of nicotinic acid mononucleotide adenylyltransferase from *Staphylococcus aureus* [26] (PDB ID: 2H29). The sequence identity between these two proteins is 35%; however, they share similar topologies, being composed of eight α-helices, a six-stranded parallel β-sheet and an additional β-strand.

The model obtained for the YP_287866 C-terminal region adopted a similar conformation to proteins belonging to the metal-dependent phosphohydrolase superfamily. These proteins possess a variety of uncharacterized domains associated with nucleotidyltransferases from bacteria, archaea and eukaryotes; YP_287866 also appears to possess one of these domain architectures. The limitation of low sequence identity (~ 25%) between YP_287866 and these proteins was circumvented by the presence of a metal-binding HD motif [32] in YP_287866. Crystal structures of HD-domain proteins have been solved for *Bacillus halodurans* (PDB ID: 2O08) and *Streptococcus agalactiae* (PDB ID: 2OGI); however, a large number of the HD-domain proteins remains uncharacterized [33].

Concerning the C-terminal region of YP_287866 (YP_287866C), the template used was the crystal structure of the putative metal-dependent phosphohydrolase from *S. agalactiae* (PDB ID: 2OGI). The resulting model consisted of an all-alpha structure formed by 13 helices.

YP_287866 is encoded by only one gene; however, it comprises two distinct domains with different functions. The complete model showed both domains linked by a disulfide bond between Cys74 and Cys275 within the N-terminal and C-terminal regions, respectively. This domain architecture was also found in another HD-domain protein fused to a nucleotidyltransferase domain [32]. Because the binding sites in both domains are not spatially superimposed, and the templates form dimers (2H29 and 2OGI), we can conclude that this architecture is likely to exist. Moreover, the model has 97.9% of its residues in preferred and allowed regions of the Ramachandran plot, indicating good stereochemical quality.

As mentioned above, some enzymes of the NAD biosynthetic and recycling pathways have not been identified in *M. hyopneumoniae*. However, based on structural information, we propose that one of the YP_287866 domains is NadD, and we also suggest that YP_287786 functions in this same metabolic pathway, thereby completing the NAD biosynthetic pathway.

The threading programs suggested the crystal structure of nicotinate phosphoribosyltransferase from *Thremoplasma acidophilum* (TmNAPRTase) [34] (PDB ID: 1YTK) as the best hit for the YP_287786 sequence. Further structural analysis suggested another homolog with a solved 3D structure, i.e., NAPRTase (EC 2.4.2.11) from *Enterococcus faecalis* (EfNAPRTase) (PDB ID: 2F7F). This enzyme catalyzes the synthesis of nicotinic acid mononucleotide (NAMN) from adenine and phosphoribosyl pyrophosphate (PRPP), regardless of the presence of ATP.

Although the sequence similarities between YP_287786 and its structural homologs TmNAPRTase and EfNAPRTase showed low overall identity (~ 25%), many residues were found conserved, among which were TmNAPRTase residues Arg224, Asp226, Glu273 and Glu292 involved in NAMN binding [34]. Two other residues also implicated in NAMN binding are found in TmNAPRTase and substituted in YP_287786, i.e., Thr179/Ser166 and Thr293/Val294. The first substitution, between amino acids having a similar physicochemical property, may not affect the function of YP_287786 because NAMN binds TmNAPRTase through a hydroxyl group.

To transfer the phosphoribosyl group, PRPP must bind to NAPRTase. Two conserved motifs, 275hSGGh279 (h stands for hydrophobic residue) and 298GVG301, are responsible for accommodating the phosphate group of PRPP. Both motifs are conserved in YP_287786 except for a glycine residue being replaced by a serine at position 277. The stereochemical quality of the YP_287786 model was verified by the Ramachandran plot calculated using PROCHECK, which showed 97% of the residues in preferred or allowed positions.

Filling gaps in *M. hyopneumoniae* pathways

The biosynthesis of flavin adenine dinucleotide (FAD) in prokaryotes involves bifunctional proteins belonging to the FAD synthetase family that catalyze both riboflavin (RF) phosphorylation and flavin mononucleotide (FMN) adenylylation. In our study, the sequence of YP_287675 showed similarities to the crystal structure of FAD synthetase (TM379) from *T. maritime* [35] (PDB ID: 1S4M) and the *in silico* model of FAD synthetase from *Corynebacterium ammoniagenes* [36] (*Ca*FADS) (PDB ID: 2X0K). Using the comparative genome tool from Genesul, we noticed that FAD synthetase was annotated in other mycoplasma genomes and YP_287675 also belongs to this cluster.

The 3D structure obtained for YP_287675 showed an overall topology similar to its template 1S4M. As expected, these proteins are folded in two domains. The N-terminal domain contains the FMN adenylylation function, catalyzing the reaction between ATP and FMN to form pyrophosphate and FAD (EC 2.7.7.2). Structurally, this domain consists of a typical nucleotide-binding fold (Rossmann fold) containing an ATP-binding site. The motif V/IXG$X_{1\text{-}2}$GXXGXXXG/A associated with the Rossmann fold and FMN binding is present in YP_287675 with a few amino acid substitutions, i.e., VX_3GGX_2AX_3GX_7A. This motif was important in assigning biological function to proteins with unknown function from fully sequenced genomes [37]. Moreover, these residues are located in conserved positions allowing substrate binding. Similarly, the residues believed to be involved in ATP-binding are conserved between YP_287675 and its template, except for Glu25 and Phe100 (replaced by aspartate and tyrosine, respectively, in 1S4M and 2X0K).

The second domain of YP_287675, the C-terminal domain, folds into a six-stranded, antiparallel β-barrel architecture, implicated in RF binding. This interaction also involves a long α-helix and a conserved histidine at position 233. RF phosphorylation by *Ca*FADS involves three important residues, Thr208, Asn210 and Asp268 [36]. With respect to sequence, none of these residues are at the same positions in YP_287675; however, the asparagine is maintained at the same structural location. Despite lacking structural information for some regions, the 3D structure of YP_287675 revealed that 96.5% of the residues are in favored and allowed regions.

The understanding of mycoplasma metabolism requires adequate annotation of its proteome. Our structure-based annotation of the proteins YP_287866, YP_287786 involved in NAD biosynthesis and YP_287675 implicated in FAD biosynthesis fills gaps in this annotation. Furthermore, proteins required in these biosynthetic pathways are being considered as antimicrobial drug targets.

Two important proteins implicated in transcription may not be absent from *M. hyopneumoniae*

The hypothetical protein YP_287559 exhibited structural similarities to the prokaryotic transcription factor NusB. NusB participates in the antitermination process, in which RNA polymerase is prevented from reading specific RNA secondary structures that usually terminate transcription. In *E. coli*, antitermination involves at least three Nus proteins: NusB, NusE (identical to the ribosomal protein S10), and NusG [38]. NusB, in association with these other proteins, is believed to bind an RNA motif, *boxA*, present in *E. coli rrn* operons. Mutations in NusB lower growth rate, which is an evidence for its role in rRNA synthesis [39]. *E. coli* has seven *rrn* operons whereas *M. tuberculosis* [40] and *M. hyopneumoniae* have only one such operon. Therefore, an efficient antitermination mechanism is particularly important in these pathogenic bacteria to ensure the expression of the entire single *rrn* operon [41]. Except for NusB, all other proteins required for efficient antitermination, such as NusA, NusG and S10, have been annotated in *M. hyopneumoniae*.

YP_287559 has only 133 residues (of 216) that align with the NusB sequence annotated in other bacterial genomes, including other species of mycoplasma. The remaining sequence (residues 1–82) possesses similarities to a transposase. As no suitable template was found to build the 3D structure of this part of the protein, only its C-terminal region was modeled.

The three dimensional structures of *E. coli* NusB [42] (PDB ID: 1EY1) and *Aquifex aeolicus* NusB [43] (PBD ID:

2JR0) derived from NMR experiments and the crystal structures of NusB from *Thermotoga maritime* [44] (PDB ID: 1TZT), *M. genitalium* [45] (PDB ID: 1Q8C), and *M. tuberculosis* [46] (PDB ID: 1EYV) were used as templates to model YP_287559.

The C-terminal portion of YP_287559 displays a topology composed of only alpha helices. Its structure can be divided into two subdomains, α1-α3 forming the N-terminal region and α4-α7 encompassing the C-terminal subdomain. In the N-terminal region, YP_287559 contains the conserved, positively charged residues Lys83, Arg84, Arg85 and Arg88, forming an arginine-rich motif with a high probability of being the RNA binding site of this protein. Also, interactions between nucleic acid bases and RNA binding proteins often involve aromatic residues essential for stacking [47]. As found in other NusB proteins, the YP_287559 sequence contains the following aromatic residues: Tyr96, Trp98, Phe101, Tyr114, Phe115, Phe127, Tyr132, Phe134, Trp147, Trp149, Phe168, Phe169, Phe176, Phe186, Phe194, Phe196, Tyr207, Tyr208, and Phe214 (Fig. 2). These amino acids located on the surface of the protein are believed to participate in recognition processes, whereas the remaining residues are probably involved in protein fold stabilization.

Previous studies have determined that NusB exists as a homodimer in *M. tuberculosis* (*mtu*NusB) [46], as a monomer in *E. coli* (*eco*NusB) [42], *M. genitalium* (*mge*NusB) [45], and *A. aeolicus* (*aq*NuB) [43], and as a monomer/dimer equilibrium with a preference for the monomeric form [44] in *Thermotoga maritima* (*tma*NusB). We searched the YP_287559 structure for amino acids important for *mtu*NusB dimerization. However, two key residues in mtuNusB, alanine and phenylalanine, are replaced by serine and tyrosine, respectively, in both *M. hyopneumoniae* and *E. coli*. In *mtu*NusB, the dimer interface overlaps the region involved in RNA binding, which may allow *mtu*NusB to remain inactive until needed for transcriptional regulation [46].

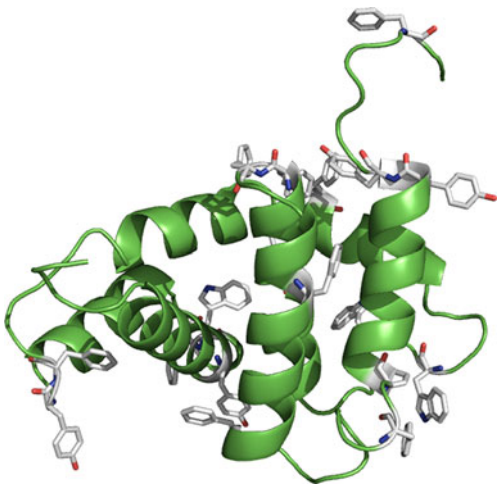

Fig. 2 The 3D structure of YP_287559. Highlighted in green are α-helices and loops; sticks represent aromatic residues likely involved in substrate recognition

We concluded that YP_287559 is composed of two domains, one similar to a transposase and the other to NusB. The Ramachandran plot analysis of the model structure from this last region showed that 95.1% of the residues are in favored and allowed regions.

The *M. hyopneumoniae* habitat is the porcine mucosal surface where amino acids, purines, and pyrimidines are acquired to compensate for the lack of important metabolic pathways. Studies suggested that, in mycoplasmas, genes involved in replication, transcription and translation are constitutively expressed in constant environments, eliminating the need for sophisticated genetic control mechanisms [1]. Moreover, *M. hyopneumoniae* has only one annotated sigma factor, RpoD [3], a key regulator of bacterial transcription initiation that is responsible for promoter recognition and melting [48]. However, the −35 regions of *M. hyopneumoniae* promoters have low sequence conservation, suggesting the presence of more than one sigma factor to respond rapidly to environmental changes.

In our structure-based analysis, we found similarities between the YP_288024 structure and the crystal structures of *Rhodobacter sphaeroides* SigE [49] (PDB ID: 2Z2S) and the flagellar Sigma-28 of *A. aeolicus* [50] (PDB ID: 1RP3). These similarities could indicate that mycoplasmas have a regulatory system not yet identified by traditional tools. Although gene expression in mycoplasma is not well characterized, recent work investigating transcriptional changes has shown that *M. hyopneumoniae* regulates its genes in response to environmental changes [51–54], and 93% of its intergenic regions are transcribed [55].

The sequence alignment of the sigma -70 family revealed the conservation of four regions, divided into subregions. Highly conserved among all members of this family are subregions two and four that compose the sigma factor binding site for the −10 and −35 promoter elements [56]. Conserved only in a highly related sigma factor, subregion one is apparently involved in an antagonistic DNA-binding activity. Subregion three is absent from YP_288024 and from extracytoplasmic function sigma factors that allow bacteria to adapt rapidly to environmental changes. Furthermore, subregion three of extracytoplasmic function sigma factors interacts with the −10 element of promoters lacking a −35 element.

The structural alignment between these proteins showed the complete lack of α-helices four and five and a portion of α-helix six corresponding to the subregion three. All the other α-helices are conserved in YP_288024, suggesting their interaction with the −10 and −35 promoter elements. This functional prediction was based on a model where

96.2% of the residues lie in the most favorable and allowed regions.

High homology to protein with unknown function

The hypothetical protein YP_287971 exhibited structural homology to YlxR from *S. pneumoniae* [57] (PDB ID: 1G2R), a small protein with unknown function, although the YlxR gene is probably in an operon with the other well-studied genes *nus*A, *inf*B, and *rbf*A. The protein encoded by *rbf*A (RbfA) binds to the 30S ribosomal subunit, perhaps promoting subunit maturation [58]. Crucial for translation initiation, IF2 (the product of *inf*B) also functions by binding the 30S subunit [59]. NusA is a highly conserved, essential elongation factor that binds RNA polymerase as part of the transcriptional antitermination complex in many organisms [60]. The YlxR-containing operon has also been studied in *E. coli* and *B. subtilis* [61]. The latter presents two additional genes (Ylx-R and Ylx-Q) between *nus*A and *inf*B; this order was not found in *E. coli* nor in *M. hyopneumoniae* wherein these genes are adjacent.

The 3D structure of YP_287971 showed a similar topology to YlxR of *S. pneumoniae*. Besides a short 3_{10}-helix, no regular secondary structure was found in the N-terminal region. The central core of the model was comprised of three antiparallel β-strands followed by two α-helices, one of which bends at Lys61. The YP_287971 sequence also possesses highly conserved residues, such as the GRGA(Y/W) motif present in the hydrophobic core together with Val10, Leu20, Leu24, Ile32, Ile47, Phe63 and Leu79. At the protein surface several positively charged residues are conserved (Arg6, Arg22, Asp27, Arg43, Lys60, Lys61 and Arg65), forming a patch typical of nucleic acid-binding proteins, as shown in Fig. 3. This region is proposed to be related in YlxR function, which may involve an RNA-binding activity found in proteins encoded by the genes in the *nus*A/*inf*B operon [57].

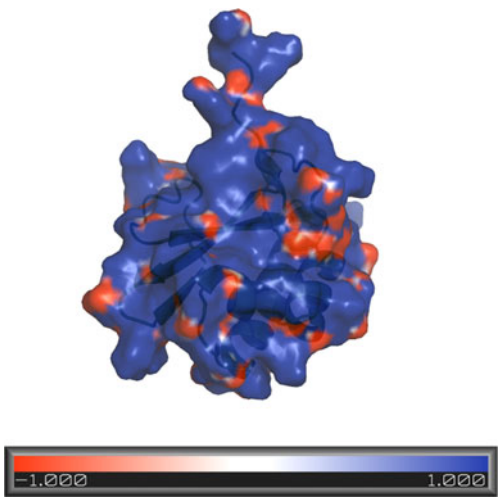

Fig. 3 Probable nucleotide binding site of YP_287971. The electrostatic potential surface distribution shows an extensive positively charged region (blue) typical of nucleic acid-binding proteins

YP_287971 is probably a member of a highly conserved family (DUF448) of unknown function, distributed in many organisms, including 14 species of mycoplasmas for which complete genome sequences are available. The stereochemical quality of YP_287971 was evaluated, resulting in 93.3% of the residues located in favored regions and 6.7% in additional allowed regions of the Ramachandran plot. Because it is of high quality and shows a significant structural resemblance to YlxR of *S. pneumoniae*, the model suggests the same function for YP_287971 and YlxR, and it will aid in the design of future experiments to verify the function.

Finally, the YP_288034 protein showed structural similarities to the crystal structure of YrdC from *E. coli* [62] (PDB ID: 1HRU). Members of the *yrd*C family code for proteins that fold into a single domain, as in the case of 1HRU, or as a domain in proteins implicated in regulation process. YP_288034 is probably an example of the latter because its alignment with *E. coli* YrdC involves only 164 amino acids out of the YP_288034 total of 287 residues. Searching for homologs within mycoplasmas, we observed that this protein clusters with a Sua5-like translation factor found in six other species. Thus, YP_288034 constitutes a two-domain protein containing a YrdC domain as found in *E. coli* and in Sua5 members such as that from *Saccharomyces cerevisiae.*

The function of *E. coli* YrdC is unknown, but its crystal structure suggested that it possesses a double-stranded RNA-binding capacity [62]. The Sua5 protein, containing an YrdC homolog domain in yeast, has been implicated in the re-initiation of translation [63]. This function is consistent with the large concave surface of Sua5; this surface has a positive electrostatic potential akin to that of the YrdC binding surface, which resembles other nucleic acid-binding proteins. The geometry of our model shows 96.6% of the residues in the most favored and additionally allowed regions of the Ramachandran plot.

Conclusions

One of the key challenges in the post-genomic era is the prediction of function for proteins annotated as hypothetical proteins. A combination of bioinformatic tools, focused not only on sequence analysis but also on structural information, guided us to suggest functions for seven hypothetical proteins in the *M. hyopneumoniae* genome. NadD, NAPRTase and FAD synthetase involved in metabolic processes; NusB and SigE in transcription; and for YrdC and YlxR, no conclusive functions were assigned; however, the results obtained helped us design rational experimental strategies for future

works. Our results suggest that this structure-based approach provides significant improvements to domain and function prediction, especially for minimal genomes having poorly annotated metabolic pathways. Mycoplasma metabolism requires an adequate annotation of its proteome, and our results fill significant gaps in this annotation. Each target protein used in this work was approached from a unique perspective, taking into account the genomic localization/organization of its open reading frame, its conserved structural features, and any biological evidence available in the literature, even if such evidence was for remote homologs. The annotation of each target required an intense effort. However, our results proved to be important for both structural and biochemical genomics.

Acknowledgments MMF thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for a PhD fellowship.

References

1. Razin S, Yogev D, Naot Y (1998) Molecular biology and pathogenicity of mycoplasmas. Microbiol Mol Biol Rev 62:1094–1156
2. Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen W-H, JaH W, Güell M, Martínez S, Bourgeois R, Kühner S, Raineri E, Letunic I, Kalinina OV, Rode M, Herrmann R, Gutiérrez-Gallego R, Russell RB, Gavin A-C, Bork P, Serrano L (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268
3. Vasconcelos ATR, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, Almeida DF, Almeida LGP, Almeida R, Alves-filho L, Assunc EN, Azevedo VAC, Brigido MM, Brocchi M, Burity HA, Camargo AA, Camargo SS, Carepo MS, Carraro DM, Castro LA, Cavalcanti G, Chemale G, Collevatti RG, Cunha CW, Dallagiovanna B, Dambro BP, Dellagostin OA, Falca C, Fantinatti-garboggini F, Felipe MSS, Fiorentin L, Franco GR, Freitas NSA, Grangeiro TB, Grisard EC, Guimara CT, Hungria M, Krieger MA, Laurino JP, Lima LFA, Lopes MI, Madeira HMF, Manfio GP, Maranha AQ, Martinkovics CT, Moreira MAM, Ramalho-neto CE, Nicola MF, Oliveira SC, Paixa RFC, Pereira M, Pereira-ferrari L, Piffer I, Pinto LS, Potrich DP, Salim ACM, Schmitt R, Schneider MPC, Schrank A, Schrank IS, Schuck AF, Seuanez HN, Silva DW, Silva R, Souza KRL, Souza RC, Staats CC, Steffens MBR, Teixeira SMR, Urmenyi TP, Vainstein MH, Zuccherato LW, Simpson AJG, Zaha A (2005) Swine and poultry pathogens: the complete genome sequences of two strains of. J Bacteriol 187:5568–5577
4. Razin S, Hayflick L (2010) Highlights of mycoplasma research–An historical perspective. Biologicals 38:183–190. doi:10.1016/j.biologicals.2009.11.008
5. Hutchison C, Montague M (2002) Mycoplasmas and the minimal genome concept. In: Herrmann R, Razin S (eds) Molecular Biology and Pathogenicity of Mycoplasmas. Kluwer, New York, pp 221–254
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
7. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(suppl 2):W5–W9. doi:10.1093/nar/gkn201
8. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36 (Database issue):D419–D425
9. Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13:121–130
10. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Bio 8:995–1005
11. Erdin S, Ward RM, Venner E, Lichtarge O (2010) Evolutionary trace annotation of protein function in the structural proteome. J Mol Biol 396:1451–1473
12. Pollack DJ (1997) Mycoplasma genes: a case for reflective annotation. Trends Microbiol 5:413–418
13. Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287:797–815. doi:10.1006/jmbi.1999.2583
14. Guo JT, Ellrott K, Chung WJ, Xu D, Passovets S, Xu Y (2004) PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. Nucleic Acids Res 32(suppl 2): W522–W525. doi:10.1093/nar/gkh414
15. Zdobnov EM, Apweiler R (2001) InterProScan - an integration plataform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848
16. Tatusov RL, Galperin MY, Da N, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36
17. Mazumder R, Vasudevan S (2008) Structure-guided comparative analysis of proteins: principles, tools, and applications for predicting function. PLoS Comput Biol 4:e1000151
18. Katoh K, Misawa K, K-i K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066
19. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2002) Comparative protein structure modeling using modeller. Curr Protoc Bioinformatics
20. Laskowski RA (2009) PDBsum new things. Nucleic Acids Res 37 (Database issue):D355–D359
21. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35(suppl 2):W407–W410. doi:10.1093/nar/gkm290
22. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18:342–348
23. Bi J, Wang H, Xie J (2011) Comparative genomics of NAD(P) biosynthesis and novel antibiotic drug targets. J Cell Physiol 226:331–340
24. Sorci L, Pan Y, Eyobo Y, Rodionova I, Huang N, Kurnasov O, Zhong S, MacKerell AD, Zhang H, Osterman AL (2009) Targeting NAD biosynthesis in bacterial pathogens: structure-based development of inhibitors of nicotinate mononucleotide adenylyltransferase NadD. Chem Biol 16:849–861
25. Lu S, Smith CD, Yang Z, Pruett PS, Nagy L, McCombs D, Delucas LJ, Brouillette WJ, Brouillette CG (2008) Structure of nicotinic acid mononucleotide adenylyltransferase from *Bacillus anthracis*. Acta Crystallogr F 64(Pt 10):893–898
26. Han S, Forman MD, Loulakis P, Rosner MH, Xie Z, Wang H, Danley DE, Yuan W, Schafer J, Xu Z (2006) Crystal structure of nicotinic acid mononucleotide adenylyltransferase from *Staphyloccocus aureus*: structural basis for NaAD iInteraction in functional dimer. J Mol Biol 360:814–825. doi:10.1016/j.jmb.2006.05.055
27. Kim MK, Kim YS, Rho SH, Im YJ, Lee JH, Kang GB, Eom SH (2003) Crystallization and preliminary X-ray crystallographic

analysis of quinolinate phosphoribosyltransferase of *Helicobacter pylori*. Acta Crystallogr D 59:1265–1266
28. Olland AM, Underwood KW, Czerwinski RM, Lo MC, Aulabaugh A, Bard J, Stahl ML, Somers WS, Sullivan FX, Chopra R (2002) Identification, characterization, and crystal structure of *Bacillus subtilis* nicotinic acid mononucleotide adenylyltransferase. J Biol Chem 277:3698–3707
29. Singh SK, Kurnasov OV, Chen B, Robinson H, Grishin NV, Osterman AL, Zhang H (2002) Crystal structure of *Haemophilus influenzae* NadR protein. A bifunctional enzyme endowed with NMN adenyltransferase and ribosylnicotinimide kinase activities. J Biol Chem 277:33291–33299
30. Yoon HJ, Kim HL, Mikami B, Suh SW (2005) Crystal structure of nicotinic acid mononucleotide adenylyltransferase from Pseudomonas aeruginosa in its Apo and substrate-complexed forms reveals a fully open conformation. J Mol Biol 351:258–265
31. Zhang H, Zhou T, Kurnasov O, Cheek S, Grishin NV, Osterman A (2002) Crystal structures of *Escherichia coli* nicotinate mononucleotide adenylyltransferase and its complex with deamido-NAD. Structure 10:69–79
32. Aravind L, Koonin EV (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases. Trends Biochem Sci 23:469–472
33. Zimmerman MD, Proudfoot M, Yakunin A, Minor W (2008) Structural insight into the mechanism of substrate specificity and catalytic activity of an HD-domain phosphohydrolase: the 5′-deoxyribonucleotidase YfbR from *Escherichia coli*. J Mol Biol 378:215–226
34. Shin DH, Oganesyan N, Jancarik J, Yokota H, Kim R, Kim S-H (2005) Crystal structure of a nicotinate phosphoribosyltransferase from *Thermoplasma acidophilum*. J Biol Chem 280:18326–18335
35. Wang W, Kim R, Yokota H, Kim SH (2005) Crystal structure of flavin binding to FAD synthetase of *Thermotoga maritima*. Proteins 58:246–248
36. Frago S, Martínez-Júlvez M, Serrano A, Medina M (2008) Structural analysis of FAD synthetase from *Corynebacterium ammoniagenes*. BMC Microbiol 8:160–175
37. Kleiger G, Eisenberg D (2002) GXXXG and GXXXA motifs stabilize FAD and NAD(P)-binding Rossmann folds through Cα–HO hydrogen bonds and van der Waals interactions. J Mol Biol 323:69–76
38. Zellars M, Squires CL (1999) Antiterminator-dependent modulation of transcription elongation rates by NusB and NusG. Mol Microbiol 32:1296–1304
39. Quan S, Zhang N, French S, Squires CL (2005) Transcriptional polarity in rRNA operons of *Escherichia coli* nusA and nusB mutant strains. J Bacteriol 187:1632–1638
40. Verma A, Sampla AK, Tyagi JS (1999) *Mycobacterium tuberculosis* rrn promoters: differential usage and growth rate-dependent control. J Bacteriol 181:4326–4333
41. Arnvig KB, Zeng S, Quan S, Papageorge A, Zhang N, Villapakkam AC, Squires CL (2008) Evolutionary comparison of ribosomal operon antitermination function. J Bacteriol 190:7251–7257
42. Altieri AS, Mazzulla MJ, Horita DA, Heath Coats R, Wingfield PT, Das A, Court DL, Andrew Byrd R (2000) The structure of the transcriptional antiterminator NusB from *Escherichia coli*. Nat Struct Biol 7:470–474
43. Das R, Loss S, Li J, Waugh DS, Tarasov S, Wingfield PT, Byrd RA, Altieri AS (2008) Structural biophysics of the NusB:NusE antitermination complex. J Mol Biol 376:705–720
44. Bonin I, Robelek R, Benecke H, Urlaub H, Bacher A, Richter G, Wahl MC (2004) Crystal structures of the antitermination factor NusB from Thermotoga maritima and implications for RNA binding. Biochem J 383:419–428
45. Liu J, Yokota H, Kim R, Kim SH (2004) A conserved hypothetical protein from *Mycoplasma genitalium* shows structural homology to Nusb proteins. Proteins 55:1082–1086
46. Gopal B, Haire LF, Cox RA, Colston MJ, Major S, Brannigan JA, Smerdon SJ, Dodson G (2000) The crystal structure of NusB from *Mycoplasma tuberculosis*. Nature 7:475–478
47. Oubridge C, Ito N, Evans PR, Teo CH, Nagai K (1994) Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature 372:432–438
48. Koo B-M, Rhodius VA, Nonaka G, deHaseth PL, Gross CA (2009) Reduced capacity of alternative σs to melt promoters ensures stringent promoter recognition. Genes Dev 23:2426–2436
49. Ea C, Greenwell R, Anthony JR, Wang S, Lim L, Das K, Sofia HJ, Donohue TJ, Sa D (2007) A conserved structural module regulates transcriptional responses to diverse stress signals in bacteria. Molecular Cell 27:793–805
50. Brown PN, Mathews MA, Joss LA, Hill CP, Blair DF (2005) Crystal structure of the flagellar rotor protein FliN from *Thermotoga maritima*. J Bacteriol 187:2890–2902
51. Madsen ML, Nettleton D, Thacker EL, Minion FC (2006) Transcriptional profiling of *Mycoplasma hyopneumoniae* during iron depletion using microarrays. Microbiology (Reading, England) 152(Pt 4):937–944
52. Oneal MJ, Schafer ER, Madsen ML, Minion FC (2008) Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to norepinephrine. Microbiology (Reading, England) 154(Pt 9):2581–2588
53. Hwang MH, Damte D, Lee JS, Gebru E, Chang ZQ, Cheng H, Jung BY, Rhee MH, Park SC (2011) *Mycoplasma hyopneumoniae* induces pro-inflammatory cytokine and nitric oxide production through NFκB and MAPK pathways in RAW264.7 cells. Veterinary Res Commun 35:21–34
54. Schafer ER, Oneal MJ, Madsen ML, Minion FC (2007) Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to hydrogen peroxide. Microbiology (Reading, England) 153(Pt 11):3785–3790
55. Gardner SW, Minion FC (2010) Detection and quantification of intergenic transcription in *Mycoplasma hyopneumoniae*. Microbiology 156(Pt 8):2305–2315
56. Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, Weinman O, Trester-Zedlitz ML, Darst SA (2002) Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Molecular Cell 9:527–539
57. Osipiuk J, Górnicki P, Maj L, Dementieva I, Laskowski R, Joachimiak A (2001) {\it Streptococcus pneumonia} YlxR at 1.35{Å} shows a putative new fold. Acta Crystallograph Sec D 57:1747–1751
58. Goto S, Kato S, Kimura T, Muto A, Himeno H (2011) RsgA releases RbfA from 30S ribosome during a late stage of ribosome biosynthesis. EMBO J 30:104–114
59. Caserta E, Tomsic J, Spurio R, Anna PCL, Gualerzi CO (2006) Translation initiation factor IF2 interacts with the 30 S ribosomal subunit via two separate binding sites. J Mol Biol 362:787–799
60. Yang X, Lewis PJ (2010) The interaction between bacterial transcription factors and RNA polymerase during the transition from initiation to elongation. Transcription 1:66–69
61. Shazand K, Tucker J, Stansmore K, Leighton T (1993) Similar organization of the nusA-infB operon in *Bacillus subtilis* and *Escherichia coli*. J Bacteriol 175:2880–2887
62. Teplova M, Tereshko V, Sanishvili R, Joachimiak A, Bushueva T, Anderson WF, Egli M (2000) The structure of the yrdC gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. Protein Sci 9:2557–2566
63. Na J, Pinto I, Hampsey M (1992) Isolation and characterization of SUA5, a novel gene required for normal growth in Saccharomyces cerevisiae. Genetics 131:791–801